CITE PUBLICATIONS

Journal of Analytical and Applied Computer Sciences
Volume: 2, Issue: 2, 2026 (July - December), Pages: 62-77
ISSN (Online): 3139-3349 | DOI: https://doi.org/10.67914/jaacs/rsa/2.2/62-77
Research Article

Using Markov chains and generalized additive model for rainfall categories forecasting

Muthanna Subhi Sulaiman, Osamah Basheer Shuku*

Dept. of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Iraq.

***Correspondence**
Osamah Basheer Shuku
drosamahannon@uomosul.edu.iq

Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International (CC BY-NC-ND 4.0).

Received: 10 August 2026 / Accepted: 25 October 2026 / Published: 28 November 2026

Background: In water resources management and flood risk reduction, accurate rainfall category forecasting is a crucial and complex climatic element due to its inherent variability in accuracy, particularly in semi-arid environments where precipitation rates are often unstable. Developing a framework for forecasting daily rainfall categories in Mosul, a semi-arid city in Iraq, is of paramount importance.

Methods: This study classifies rainfall data into four categories based on rainfall quantities: dry, light, moderate, and heavy. The dominance of dry days (approximately 86%) in Mosul creates a data class imbalance. To overcome this, class weights and decision threshold adjustments were employed to mitigate the dominance of the prevailing category. The study compares the traditional probabilistic model, First-Order Markov Chain (FOMC) model, with Generalized Additive Model (GAM) in capturing non-linear and seasonal relationships. In this study, the AR concept is adopted in the GAM environment that allows multiple temporal lags to be incorporated. Thus, the temporal memory of rainfall and its categories can be represented. This surpasses the limitations of higher-order MC models and the strict stationarity assumptions of ARIMA models.

Results: The results show that the GAM model outperforms the MC in accuracy and probabilistic evaluation, indicating MC 's failure to predict rare events due to its near-zero probabilities. Furthermore, the enhanced model (GAM with Lag 1 - 4) demonstrates superiority over the simpler GAM model across all measures.

Conclusion: Therefore, it is possible to conclude that combining the concept of autoregression with GAM provides an advantage by being a flexible and interpretable predictive tool that outperforms other models, especially in semi-arid environments, and also provides a solid foundation for environmental risk management.

Keywords: *Rainfall categories, Generalized Additive Model (GAM), Markov chain, Forecasting*

Introduction

The random nature of rainfall and its dependence on nonlinear interactions between atmospheric variables due to its variability make it one of the most complex phenomena. In semi-arid environments, rainfall is the primary source of surface and groundwater resources and the cause of flood and drought risks. The city of Mosul in northern Iraq is characterized by significant interannual fluctuations in rainfall rates, with a clear concentration of rain in the cooler months (November-April) and dryness in the warmer months. CM is a traditional tool used to model future conditions based on

current conditions, making it suitable for analyzing wet-dry transitions. In recent studies, such as (Chowdhury, Roy, & Saha, 2025), Markov models were compared with machine learning techniques to analyze rainfall patterns. Results revealed that MC models can retain statistical properties of drought and rainfall continuity, while machine learning techniques performed better in daily forecasting. FOMC modeling assumes stationarity through time and disregards seasonality of rainfall in semi-arid environments. Thus, more flexible models have been used. Generalized ensemble models (GAMs) have been found to be better in representing nonlinear relationships between explanatory and dependent variables in recent studies, having the flexibility to include spline functions to capture seasonality, especially in the statistical modeling of daily rainfall (Chhatre, O'Donnell, & Ravishanker, 2026; Yusif, 2024). Most of the studies carried out on rainfall forecasting in Nineveh and Mosul governorates were only interested in forecasting magnitude using machine learning algorithms or traditional time series models. However, there is very little research on rainfall category forecasting in Mosul using interpretable association models. In this work, we attempt to fill this gap by proposing an integrated framework combining probabilistic MC modeling and flexibility of GAM models, especially after aligning them with the concept of AR. In place of the complex box-Jenkins methodology and its assumptions on stability for building an AR model, the GAM model uses multiple temporal lags as explanatory input variables. This hybridization makes it possible for the model to take into account the temporal memory of rainfall with interpretability and without computational complexities. The proposed framework also tackles the issue of class imbalance in semi-arid climates, where drought days are the dominant pattern, by the simultaneous application of class weights and adjusted decision thresholds.

Materials and Methods

Theoretical framework

The theoretical framework of this study gives a systematic approach to both the probabilistic Markov modeling and ensemble models. The framework is grounded in three key theoretical bases as follows:

First: Probabilistic Categorization: Instead of a continuous variable with high dispersion and extreme values, rainfall amounts are transformed into a categorical variable. The class boundaries are determined using percentiles from the data to represent the balanced classes as well as possible, transforming the prediction problem into a classification problem, which leads to more efficient probabilistic modeling.

Second: Traditional probabilistic modeling with Markov chains: This approach uses FOMC which assumes that the day's weather conditions depend only on the previous day's weather conditions. The transition matrix is a measure of the transition between the states with probability of the transitions. Assumptions of stationarity and the neglect of seasonality may render the Markov chain unable to adapt to climatic variations across seasons and years.

Third: Modeling Using GAM: To improve the predictive results of CM and overcome its limitations, a GAM model will be employed. This model allows for the representation of non-linear relationships between explanatory and dependent variables. It relies on spline functions to capture seasonality and estimate the probabilities of rainfall classes sequentially across different periods. Time is used for greater alignment with climatic realities. This framework incorporates two methodological innovations. The first involves using the concept of autoregression (AR) instead of constructing a stationarity-dependent ARIMA model. This is achieved by using multiple temporal lags as explanatory input variables for the GAM model. This hybridization is akin to using a higher-order Markov, allowing the model to capture the temporal components of rainfall that is, the influence of previous days' rainfall on the current day. The second innovation is the class imbalance problem, as drought days account for 86% of semi-arid days. Traditional models tend to predict the majority class for the highest accuracy, which is an illusion due to a phenomenon called the precision paradox. Thus, class weights will be used in the developed model (GAM with Lag 1-4). Decision thresholds will also be altered. To ensure fair comparison, thresholds were applied to all models (CM, Simple GAM and Advanced GAM) to break the drought dominance and improve rainfall forecasts. The steps of the systematic research are shown in Figure (1) below. First, the data is collected and cleaned on a daily basis; then, it is segmented into seasons and converted into categories; next, the CM, Simple GAM, and Advanced GAM models are constructed; finally, the Log-Loss, F1-Score, and Sensitivity metrics are used to evaluate the models' performance.

Markov Chain (CM)

CM is a probabilistic model used to model systems that change state over time, where the future state depends solely on the current state, regardless of previous states. This principle is known in statistical literature as the Markov Property or

absence of temporal memory (Hajani & Sarma, 2023). If the random variable X_t represents the state of the system at time t, then it is possible to express the Markov property as in the following equation:

$$P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \dots, X_0 = i_0) = P(X_{t+1} = j | X_t = i)$$

When i and j represent states of a state space.
Despite its simplicity and effectiveness in modeling random phenomena, CM suffers from practical limitations for climatic phenomena, including that it assumes that the probabilities of transitions in the series are constant over time, which is inconsistent with climatic systems characterized by seasonality. It also does not take into account that the probabilities of rainfall vary drastically between the seasons. Furthermore, if no transition occurs, the corresponding probability of transition will be zero, which leads to a failure to predict rare events (Hyndman & Grunwald, 2000).

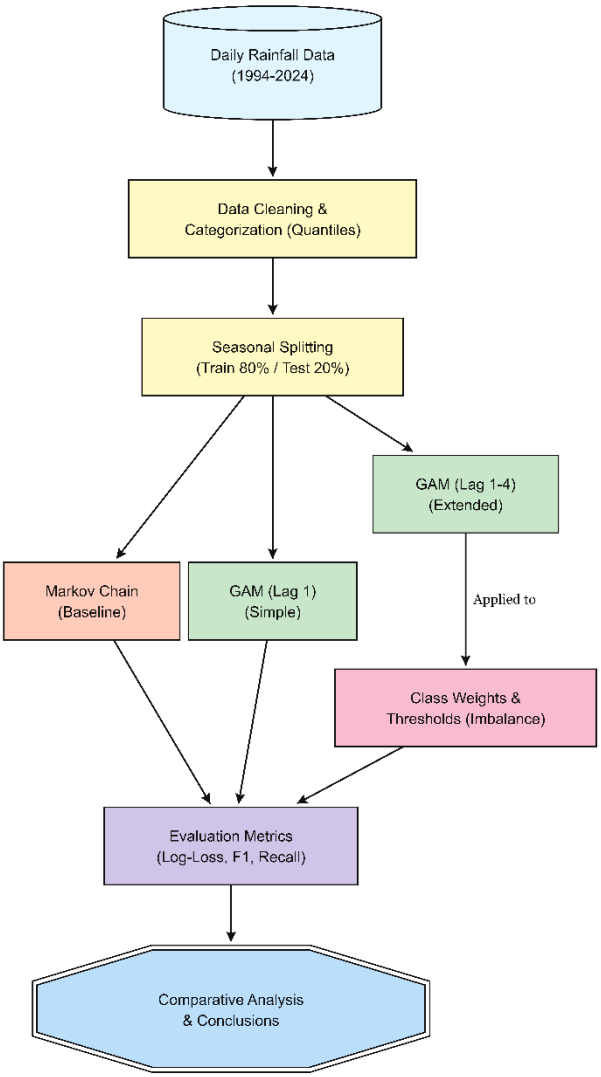


Figure 1. Research methodology sequence diagram

Transition probability matrix

For discrete states, the state space S can be defined as the set of possible states. The properties and probabilities of transitions between these states are described using the Transition Probability Matrix, denoted by P. It is a matrix of size $K \times K$, where K represents the number of states, and its elements can be defined as p_{ij} (Hajani & Sarma, 2023).

$$p_{ij} = P(X_{t+1} = j | X_t = i)$$

Since p_{ij} represents the probability of transitioning from state i at time t to state j at time t+1, the transition matrix has mathematical properties including that all elements are non-negative: $p_{ij} \geq 0$ for all i and j. The sum of each row is equal to one: $\sum_{j=1}^{K} p_{ij} = 1$ for all i. Therefore, the matrix takes the following form:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix}$$

Estimating the Transition Matrix: Transition probabilities, p_{ij}, are often unknown and must be estimated from historical data. The Relative Frequency Approach, the most commonly used method for estimating these probabilities, is equivalent to Maximum Likelihood Estimation (MLE). Therefore, $\hat{p}_{ij}$ will be calculated as follows (Ohaegbulem & Nwobi, 2026).

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{K} n_{ij}}$$

n_{ij} represents the number of times the system transitions from state i to state j in time t+1, while the denominator includes the total number of times the system leaves state i**.**

Predicting rainfall categories using CM

After calculating the transition matrix $\widehat{P}$, it can be used to predict the following day's conditions. To predict rainfall categories, the Maximum a Posterior Estimator (MAP) or the Argmax Rule (the rule with the highest probability of a transition) will be used. According to this rule, the category j^* with the highest probability of a transition is predicted, as follows:

$$j^* = argmax_{j \in S} \hat{p}_{ij}$$

Therefore, the model will predict the category with the highest probability of occurrence in the next day's scenario based on the current day's scenario (Chowdhury et al., 2025).

The Generalized Additive Model (GAM)

The Generalized Additive Model (GAM) is an extension of the Generalized Linear Models (GLM) and was developed by (Hastie & Tibshirani, 1986) by replacing the linear relationship between the independent and dependent variables with spline functions, which allows the model to adapt to nonlinear patterns in the data. The GAM can be expressed as in the following equation (Wood, 2017).

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$$

Where g(.) represents the link function, μ represents the mean of the dependent variable, β_0 is the intercept, $f_{i(x_i)}$ represents the smooth functions of the independent variables, and ε is the random error limit. The effectiveness of GAM depends on choosing the appropriate base to represent the smooth functions f_i. The cubic spline base is commonly used in climate modeling when the range of the independent variable is divided into sub-knots with cubic polynomials fitted to each sub-knot. The smooth function can be expressed as a linear form of the base functions as follows:

$$f(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$

When $b_{k(x)}$ represents the basis functions and β_k are the parameters. To control the degree of smoothness of the function, and to prevent overfitting in seasonal data modeling, a penalty is introduced on the coefficients, and a cyclic basis is used to ensure that the smooth function is continuous at the end and beginning of the year to conform to the seasonal and cyclical nature of the climate (Hyndman & Grunwald, 2000; Shukur, 2021). A multinomial family model will be used because the dependent variable in this study is a categorical variable consisting of four categories: dry, light, moderate, and heavy. Since K represents the number of categories, K-1 equations will be constructed, each comparing a specific category to a reference category. The probability that day t belongs to category j can be represented by the following equation:

$$P(Y_t = j|x_t) = \frac{exp\left(\eta_j(x_t)\right)}{1 + \sum_{l=1}^{K-1} exp\left(\eta_l(x_t)\right)}$$

When $\eta_{j(x_t)}$ represents the linear explanatory variable of class j, and x_t is the vector of the independent variables on day t, then $\eta_0 = 0$ (Wood, Pya, & Säfken, 2016).

To overcome the limitations imposed by the traditional Markov chain, the concept of autoregression (AR) will be employed by introducing multiple temporal lags as covariates into the GAM model. This will allow the model to capture the temporal patterns of rainfall as well as the cumulative effect of days in the past on the probability of rain occurring on the present day. The explanatory variables in the developed model (Lag 1-4) can then be expressed as follows:

$$\eta_j(x_t) = \beta_{0j} + f_{1j}(DY) + \sum_{k=1}^{4} \beta_{kj}(Lag_k)$$

DY represents the day of the year (for seasonality), and Lag_k represents the category of the previous day, given that k=1,2,3,4 and β_{0j} and β_{kj} represent the parameters that are estimated for each category j.

This procedure is equivalent to raising the CM rank to higher ranks without excessive computational complexities to capture seasonal rain persistence more flexibly (Anfossi, Cappa, Von Hardenberg, & Provenzale, 2002; Hasan & Shukur, 2025). As for the problem of the dominance of the dry, prevailing category by approximately 86%, it can be addressed by using two complementary procedures. This is done by giving higher weights to the rare rainfall categories and lower weights to the dominant category while training the model to penalize errors in the rare categories more severely. Dedicated decision thresholds are also used, which are modified for each category, in order to convert probabilities into classification decisions. At that point, the modified prediction rule is defined as follows:

$$\hat{y}_t = argma\, x_{j\in\{1,2,3\}}\left(P(Y_t = j|x_t) - \tau_j\right)$$

When τ_j represents the threshold set for class j, the class that exceeds its threshold by the largest difference will be selected, and thus rare events are discovered without being dominated by the most frequent class (Vu, Nguyen, Nguyen, Thi, & Nguyen, 2025). If GAM is compared to a Markov chain, it is distinguished by its ability to capture complex relationships and patterns between climatic variables and rainfall probabilities. Furthermore, by using periodic smooth functions, it will realistically take seasonal changes into account, and it will allow understanding the effect of each explanatory variable on the dependent variable through smooth functions (Hyndman & Grunwald, 2000).

Framework of AR-GAM hybridization

The most significant contribution of this research is the integration of the autoregressive (AR) concept within the Generalized Associative Model (GAM) environment. This integration aims to construct an algorithm capable of capturing the temporal patterns of rainfall classes and overcoming the rigid assumptions and conditions imposed by ARIMA and higher-order Markov models. The main advantages of this AR-GAM hybridization can be summarized as follows:

Overcoming the limitations of higher-order Markov models, which require constructing a $K^4(4^4 = 256)$ transition matrix, leading to dimensionality explosions. Using time lags as collectible input variables instead of discrete states provides a suitable solution to circumvent these limitations. The non-linear flexibility allows for the continued modeling of seasonal and cyclical variations using spline functions. Bypassing the stationarity assumption inherent in ARIMA modeling and its complexities, as well as the variability of seasonal characteristics, is achieved by considering time lags as input variables in the GAM to adapt to seasonal data changes over time. The hybridization algorithm can be implemented in a series of steps, as illustrated below and shown in the flowchart in Figure 2.

Step 1: Construct the lag structure from the original series of rainfall categories, considering the weather vector for the previous day (Lag 1), two days prior (Lag 2), and so on up to Lag 4. This results in the variables $X_{t-1}, X_{t-2}, X_{t-3}, and\ X_{t-4}$.

Step 2: Factor encoding using factor variables allows the GAM model to estimate a separate coefficient for each previous state (Dry, Light, Moderate, Heavy).

Step 3: Integration with Seasonality: This involves incorporating the variable (DY) as a cyclic smooth function to capture seasonal patterns.

Step 4: Constructing the Additive Model. The linear explanatory variable for class j is the sum of the following effects.

$$\eta_j(x_t) = \beta_{0j} + f_{1j}(DY) + \sum_{k=1}^{4} \beta_{kj}(Lag_k)$$

Step 5: Imbalance handling: We handle the imbalance by using class weights in the training and then adjusting the thresholds in the prediction.

Step 6: Forecast and Evaluation. We forecast the test data using the metrics of Log-Loss and F1-Score. This newly adjusted threshold is then used for classification decisions from the data.

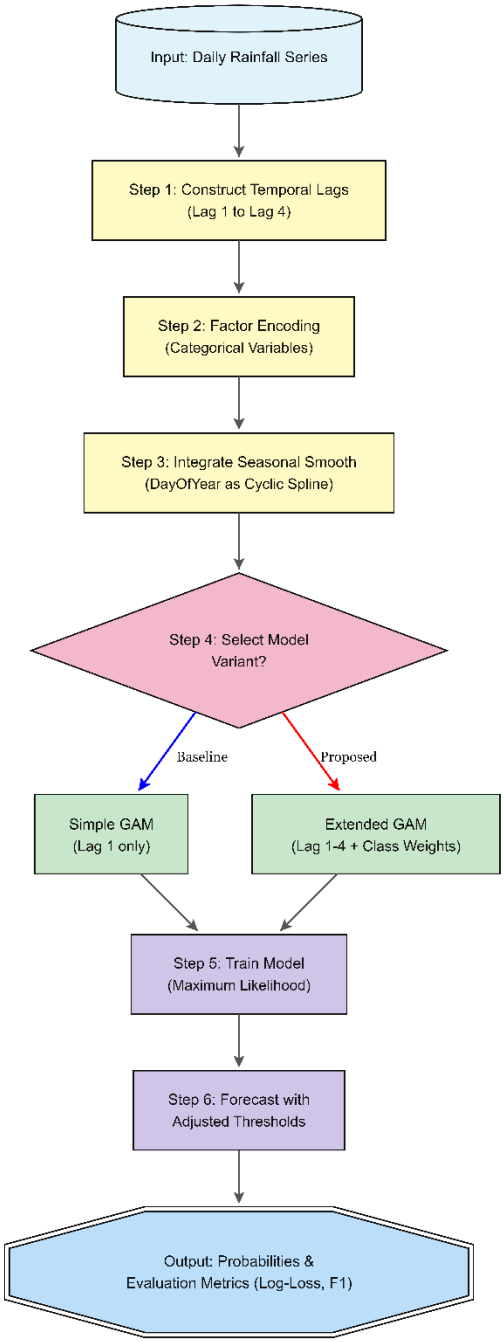


Figure 2. Flowchart of the theoretical hybridization algorithm

Performance evaluation metrics

In the climatic environment of this study there is a large class imbalance, with the dry class being the dominant class at approximately 86%, so using only the precision metric is misleading and can lead to what is known as the "precision paradox." Thus, the study employs a set of statistical metrics to assess performance from multiple angles, with particular emphasis on the model's ability to predict rare classes (rainfall).

Confusion matrix

The confusion matrix is the basis for classification metrics. This is a table showing the number of correct and incorrect predictions for the classes, comparing the actual values with the classified values. The matrix elements for class j are defined as:

- True Positive (TP): The number of days when the model's classifications were in class j and they were indeed in class j.
- True Negative (TN): number of days in which a model's classifications were not of class j and were in fact not of class j.
- False Positive (FP): The number of days on which the model's classifications were of class j but were in fact of another class (Type 1 false).
- False Negative (FN): The number of days on which the model's classifications were not of class j but were in fact of class j (Type 2 false). The matrix can be formulated as follows (Grandini, Bagli, & Visani, 2020).

$$C = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}, \qquad \text{where } N_{total} = \sum_{i,j} C_{i,j} = TP + FN + FP + TN$$

Precision and sensitivity

Precision is measured as the proportion of correct classifications of category j out of the total number of classifications the model assigns to category j. It is expressed as follows:

$$\text{Precision}_j = \frac{TP_j}{TP_j + FP_j}$$

This differs from Accuracy, which refers to the overall accuracy of the model as a whole. Accuracy is highly misleading when data are unbalanced and is not used in multi-category problems (such as classifying rainfall categories, which in our study are divided into four main categories). Accuracy is expressed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity, or recall, is measured as the proportion of days correctly classified as category j out of the total number of actual days belonging to category j. It is expressed as follows:

$$\text{Sensitivity}_j = \frac{TP_j}{TP_j + FN_j}$$

Sensitivity is a very important measure in the context of rainfall classifications, as it reflects the model's ability to detect actual rainfall events (False Negatives) (Sokolova & Lapalme, 2009).

F1 score

Accuracy is often inversely related to sensitivity; therefore, the F1 score is used as a harmonic mean to provide a balanced measure of classification performance. It is defined as follows:

$$F1_j = 2 \times \frac{\text{Precision}_j \times \text{Sensitivity}_j}{\text{Precision}_j + \text{Sensitivity}_j}$$

The F1 score is the most suitable measure for unbalanced data because it gives equal weight to accuracy and sensitivity, especially in cases of dominant class. It is calculated as a macro-F1 score, which is the arithmetic mean of the F1 scores for all classes to measure the overall performance of the model (Sokolova & Lapalme, 2009).

Log-Loss

The log-loss is the most accurate measure of the quality of the resulting probabilities for categorical classifications. It measures how severely a model is penalized if it produces low probabilities for events that actually occur. It is defined as follows:

$$Log - Loss = -(1/N)\sum(i=1)\text{^}N\text{▒}\sum(j=1)^{K} y_{ij}\, log\big(p_{ij}\big)$$

where N is the number of days, K is the number of categories (4), y_{ij} is 1 if day i actually belongs to category j, and 0 otherwise, and p_{ij} is the expected probability of category j on day i. A lower Log-Loss value indicates better performance because a model that produces near-zero probabilities for an event that actually occurs will be severely penalized (Gneiting & Raftery, 2007).

Area under the curve (ROC-AUC)

The Receiver Operating Characteristics (ROC) curve is used to evaluate a model's ability to distinguish and classify categories across possible thresholds. This is done by measuring the Area Under the Curve (AUC), which ranges from 0.5 to 1. The closer it is to 1, the more optimal the model. AUC can represent the probability that the model will assign a higher score to a positive random sample compared to a negative random sample. It is expressed as follows:

$$\mathrm{AUC} = \mathrm{P}\big(\mathrm{S}_{\mathrm{positive}} > \mathrm{S}_{\mathrm{negative}}\big)$$

$S_{positive}$ represents the probability of a positive sample, while $S_{negative}$ represents the probability of a negative sample. AUC can be practically calculated using the Wilcoxon-Mann-Whitney Utility (WMU) statistic, as shown in the following equation:

$$AUC = \frac{(\sum R_i) - \frac{n_{pos}(n_{pos}+1)}{2}}{n_{pos} \times n_{neg}}$$

When n_{pos} represents the number of positive cases, n_neg represents the number of negative cases, and R_i represents the rank of the positive case i when the scores are arranged in ascending order.

In accordance with the data of this multi-group study, the AUC will be used and calculated for each group j separately, considering that group as positive and the rest as negative. Then, the overall mean will be calculated as in the following equation:

$$AUC_{macro} = \left(\frac{1}{K}\right)\sum_{j=1}^{K} AUC_j$$

When K represents the number of groups (Fawcett, 2006; Hand & Till, 2001).

Result and Discussion

This section will review the detailed results obtained from the proposed models, starting with a comprehensive description of the daily rainfall data at the Mosul station and how it is categorized and divided into time periods. The performance of the three models (CM, GAM, and the enhanced GAM) will also be reviewed. The results will be interpreted in light of the climatic characteristics of the semi-arid environment in Mosul, and the probabilistic and classificatory performance of the statistical models will be compared, along with the suitability of each model for predicting and classifying rainfall categories. We used the R programming language to extract the data programmatically.

The climatic characterization of the data

In this section, the daily rainfall data of Mosul station will be analyzed, some descriptive metrics and statistics will be calculated, and the climatic characteristics of the study area will be reviewed, with the problem of dispersion and imbalance in the data. Moreover, it will discuss how to transform continuous quantities into discrete categories, and how to split the data into training and testing periods, considering the boundaries of the hydrological seasons, so as to ensure a sound analysis that reflects the environmental reality of Mosul. Rainfall in Mosul is characterized by high variability and temporal irregularity due to the region being affected by low-pressure systems during the winter and spring months. The rainy season in Mosul is concentrated between November and April, with peak rainfall occurring in January, February, and March, while the summer months are dry. The data used in this study were obtained from the Iraqi Meteorological Organization and Seismological Monitoring as unpublished data for Mosul Station (lat 36.19 o N, long 43.09 o E, and alt 223 m above sea level). This data was extracted from the doctoral dissertation of (K. R. Nafea, 2026) and the research derived from it (K. Nafea & Al-Dujaili, 2026). The daily rainfall time series in this study covers the period from January 1, 1994 to December 31, 2024, a total of 11,323 days (Al-Nassar, Pelegrí, & Masciángioli, 2026). This daily time series exhibits statistical characteristics due to the semi-arid climate, with a tendency towards zero values for dry season days and extreme values for days of heavy rainfall. Table (1) illustrates some descriptive statistics for the rainfall time series in Mosul for the period (1994-2024).

Table 1. Descriptive statistics of daily rainfall data (Mosul station, 1994-2024)

Statistic	Value
Total Observations (Days)	11,323
Total Rainfall (mm)	8,960.8
Mean (mm/day)	0.79
Standard Deviation	3.08
Median (mm/day)	0.0
Maximum (mm/day)	111.0
Skewness	10.25
Dry Days (0 mm)	8,081

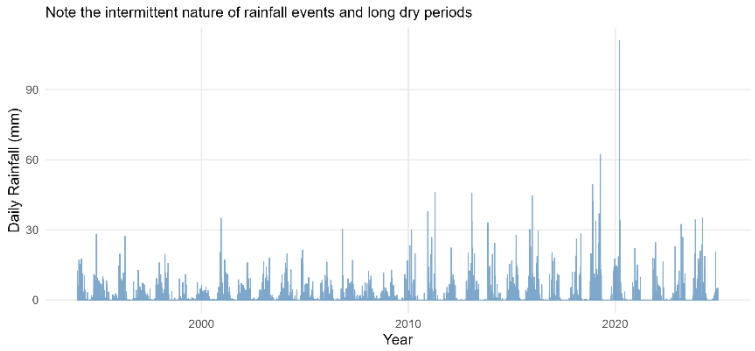


Figure 3. Daily rainfall time series for Mosul city (1994-2024).

Figure 3 shows the time series of daily rainfall amounts at Mosul station for 31 years. The graph shows clearly that the rainfall is intermittent, with periods of drought followed by short-term precipitation events, thus confirming a semi-arid climate. The series is confirmed to be highly positively skewed, with most of the rainfall being at low values close to zero, while only rare extreme peaks reach 111 mm/day (Table 1). This demonstrates that heavy rainfall events are rare. This explains the difficulty of correctly predicting continuous amounts, and the need to group continuous values to be able to catch these rare events.

Figure 4 presents interannual variability of total rainfall. The impact of the climate variability varies, resulting in a large difference between years. In some years there is a great deal of rain, while in other years there is less than the average total for the year. Hence, predictive models must incorporate sequential changes in time series using more flexible models such as GAM. Due to this climatic nature, the continuous variable was transformed into a categorical variable with four categories whose boundaries were defined by the quartiles obtained from the data as follows: Category 0 (Dry): 0.0 - 0.9 mm Category 1 (Light): 0.9 - 2.0 mm Category 2 (Moderate): 2.0 - 4.71 mm Category 3 (Heavy): greater than 4.71 mm. Table 2 shows the distribution of classes, where the dominance of the dry class is observed and the other three rainy classes are distributed in very close proportions, which indicates the imbalance of the data (Class Imbalance), which is the main problem that justifies the use of class weights.

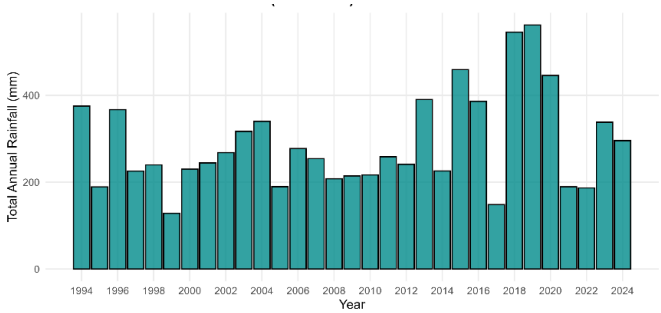


Figure 4. Annual total rainfall in Mosul city (1994-2024).

Table 2. Distribution of rainfall categories in training set

Category	Frequency (Days)	Percentage (%)
Dry (0)	7,845	85.92%
Light (1)	435	4.76%
Moderate (2)	414	4.53%
Heavy (3)	437	4.79%
Total	9,131	100%

The data were divided into two sets, a training set and a test set, according to the hydrological seasons, the rainy season being from November 1st to April 30th. The training set consisted of seasons from 1994 to 2019, which totaled 25 seasons and 9,131 days or approximately 80% of the data. The test set consisted of the seasons from 2019-2020 to 2024-2025, in total 6 seasons, 1,888 days, which is approximately 20% of the data. The summer months are characterized by almost no rainfall at all, but the time series did not contain any dry days or seasons with zero values. This is because the removal of dry days or dry months from the time series would break the chronological order and would not allow the CM to calculate the actual transition probabilities between conditions. This would eliminate the true characteristics of the data and create a misleading distinction between dry and rainy days. Transition probabilities would be misleading, correctly, and most dry days in semi-arid climate would be excluded. Thus, the dry days keeping will result in a balanced transition matrix that reflects the true climatic reality of Mosul (K. Nafea & Al-Dujaili, 2026).

Practical framework and algorithm

The implementation process will follow these sequential steps:

Step 1: Load the daily time series (11323 days) in the R language environment and set the boundaries of the four classes according to the quartiles resulted from the training data. The continuous rainfall amounts are categorized into four classes: Dry, Light, Moderate and Heavy.

Step 2: Temporal division into ~ 2 training periods (80%) and a test period (20%).

Step 3: Model construction includes three models: (i) a Markov chain to estimate the probability transition matrix from the training data, (ii) a simple GAM using smoothed seasonality functions and (iii) an extended GAM with four-time lags (Lag 1-4) and class weights to account for imbalances.

Step 4: The probabilities of classes in the test data were computed using the trained models. Decision thresholds were adjusted to convert probabilities into classifications and to solve the drought class dominance problem.
Step 5: Probabilistic performance indicators (Log-Loss) and metrics (Precision, Sensitivity, F1-Score, ROC-AUC) were used to evaluate the classification performance of the three models. Also, confusion matrices were calculated.

Step 6: The results were discussed in the light of the climatic characteristics of Mosul, comparisons were made and reasons for the superiority and weakness of the classification performance were clarified. The framework and the used algorithm are shown in Figure 5 as follows.

Log-Loss results

Log-Loss is the best metric to evaluate the quality of probabilistic forecasts. The Log-Loss calculation for the three models on the test data is shown in Table 3 below.

Table 3. Comparison of log-loss of forecasting models on testing set

Model	Log-Loss
Markov Chain	20.8633
Simple GAM (Lag 1)	0.3856
Extended GAM (Lag 1-4)	0.3851

The results, as indicated by Table 3, imply the following:

1. Statistically the CM failed because of a very high Log-Loss value. This is because of the fixed transition matrix used by CM with almost-zero probabilities for some classes in predicting a rainy day, which leads to an infinite penalty and shows its unsuitability for probabilistic classification on rare datasets.
2. The GAM models showed better performance by reducing the Log-Loss value to 0.3856 in Simple GAM model and further to 0.3851 in the improved model. This means that the model gives high probabilities to true events and low probabilities to false events, which means the probabilistic classification is realistic and balanced. The difference is small but it is practically meaningful. The improved model with four-time lags (Lag 1-4) and class weights showed a slight decrease in log-loss.

Classification metrics: Sensitivity, Precision, and F-1 score

Table 4. The classification measures for the three models on the test data after applying the adjusted thresholds are as follows.

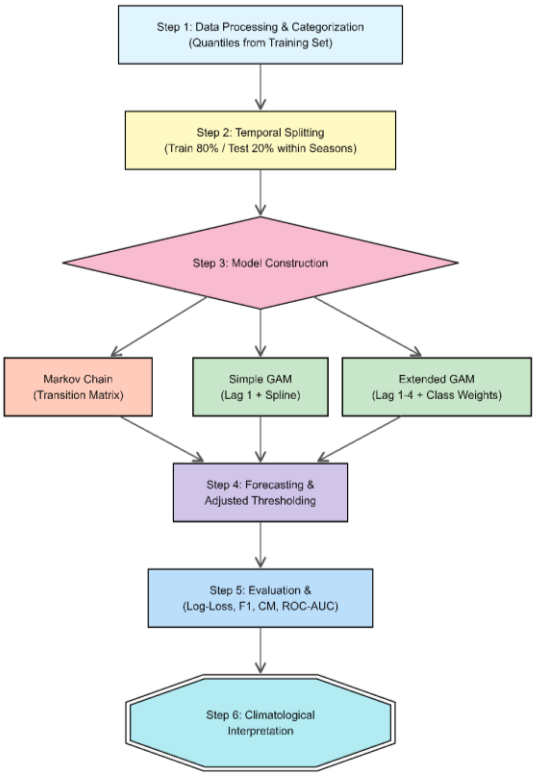


Figure 5. Flow chart of the practical framework and algorithmic sequence followed in the study.

Table 4. Classification metrics (precision, sensitivity, f1-score) on test data.

Model	Category	Precision	Sensitivity	F1-Score
Markov Chain	Light (1)	0.026	1.000	0.052
	Moderate (2)	0.000	0.000	0.000
	Heavy (3)	0.000	0.000	0.000
Simple GAM	Light (1)	0.000	0.000	0.000
	Moderate (2)	0.000	0.000	0.000
	Heavy (3)	0.222	0.253	0.236
Extended GAM	Light (1)	0.000	0.000	0.000
	Moderate (2)	0.000	0.000	0.000
	Heavy (3)	0.221	0.263	0.240

From Table 4 above, the results can be interpreted as follows:

1. As seen in the table, CM obtained an ideal sensitivity value for the Light category, which means that CM was able to predict all the days of light rain. But it was very inaccurate, and classified almost all days as light rain, even days that were dry, so it had a lot of false positives. Thus, CM relied on the padding of the rare category to avoid this penalty, and totally failed to detect and classify the Medium and Heavy categories.
2. The improved GAM model performed best in terms of accuracy and sensitivity for the rare category under Heavy rain. Thus, the inclusion of time lags (Lag 1-4) and class weights allowed the model to improve its classification ability in detecting heavy rain while maintaining the necessary accuracy.
3. The extended GAM did not predict light and moderate rainfall categories, as its F1 values were zero, same as the other two models. This is because of the hydrological nature of rainfall series in semi-arid climate and the similarity and convergence of rainfall amounts (1–4 mm) with dry days, making it impossible to differentiate between them and drought using time lags only.
4. The F1 score confirms that the extended GAM model is the best in terms of prediction for rare heavy rainfall categories while the CM model gave misleading results and the simple GAM model was the least efficient in detecting rare categories.

Confusion matrices

Confusion matrices were calculated for the training and the test period for each model in raw and normalized form. The values of TP, FP, FN and TN were also calculated for each category. The emphasis was on the confusion matrices for the testing period to demonstrate the true classification ability of generalization to new data. Table 5: Normalized Confusion Matrix of CM model on the test data.

Table 5. Normalized confusion matrix for cm in the (testing)

Actual \ Predicted	Dry (0)	Light (1)	Moderate (2)	Heavy (3)	TP	FP	FN	TN
Dry (0)	100%	0%	0%	0%	1,691	196	0	0
Light (1)	100%	0%	0%	0%	0	0	50	1,837
Moderate (2)	100%	0%	0%	0%	0	0	51	1,836
Heavy (3)	100%	0%	0%	0%	0	0	95	1,792

Table 5 clearly shows the complete failure of the CM model to predict any rainfall category, as all days for all categories were classified as dry. This demonstrates the Precision Paradox and its practical uselessness.

Table 6 shows the normalized ambiguity matrix of the simple GAM model on the test data.

Table 6. Normalized confusion matrix for simple gam in the (testing)

Actual \ Predicted	Dry (0)	Light (1)	Moderate (2)	Heavy (3)	TP	FP	FN	TN
Dry (0)	96.3%	0%	0%	3.7%	1,628	151	63	45
Light (1)	84.0%	0%	0%	16.0%	0	0	50	1,837
Moderate (2)	74.5%	0%	0%	25.5%	0	0	51	1,836
Heavy (3)	74.7%	0%	0%	25.3%	24	84	71	1,708

Table 6 clearly shows that the simple model demonstrated some classificatory ability in detecting heavy rainfall, but none of the light and medium categories were detected, and all were classified as dry days. This indicates its need to capture subtle seasonal variations in the intensity of light and medium rainfall.

Table 7 presents the normalized ambiguity matrix of the improved model (Lag 1-4 + Class Weights) on the test data.

Table 7. Normalized Confusion Matrix for Extended GAM in the (Testing)

Actual \ Predicted	Dry (0)	Light (1)	Moderate (2)	Heavy (3)	TP	FP	FN	TN
Dry (0)	96.0%	0%	0%	4.0%	1,624	150	67	46
Light (1)	84.0%	0%	0%	16.0%	0	0	50	1,837
Moderate (2)	74.5%	0%	0%	25.5%	0	0	51	1,836
Heavy (3)	73.7%	0%	0%	26.3%	25	88	70	1,704

Table 7 clearly shows that the extended GAM model demonstrates a slight improvement in detecting heavy rainfall, with the detection rate increasing by one day compared to the simple GAM model. However, light and medium rainfall remains undetectable.

Based on the results presented in the preceding figures and tables, it is evident that CM failed to detect rainfall categories for all data points, being dominated by drought prediction. Furthermore, both the simple and extended GAM models successfully detected over a quarter of the heavy rainfall days but failed to detect light and medium rainfall. This confirms that predicting heavy rainfall is the most effective approach, while classifying light and medium rainfall days presented a significant challenge in the semi-arid climate of Mosul.

Table 8 shows the ROC-AUC values for the extended GAM model for rainfall categories in the test data.

Table 8. ROC-AUC for rainfall categories for extended gam in the (testing)

Category	ROC-AUC
Light (1)	0.501
Moderate (2)	0.512
Heavy (3)	0.620

Table 8 clearly shows a significant weakness in the discriminatory properties of the light and moderate categories. The AUC for both light and moderate is approximately 0.5, clearly indicating that the model is almost unable to distinguish between light and moderate days and dry days. This confirms the previously mentioned classification results and confirms that light and moderate rainfall is hydrologically similar to the dry day category, and the four-time lags failed to differentiate between them. The heavy category, on the other hand, recorded a higher AUC of approximately 0.6, remaining within the moderate discrimination range. However, this reinforces the model's superior ability to discriminate against rare events. Heavy rainfall exhibits distinctive behavior in time memory, enabling the model to capture some seasonal patterns. Overall, the results confirm that classifying rainfall categories in semi-arid climates is complex and challenging. Figure 6 includes four curves derived from extended GAM, illustrating how the probability of each category changes across the days of the year (1-365).

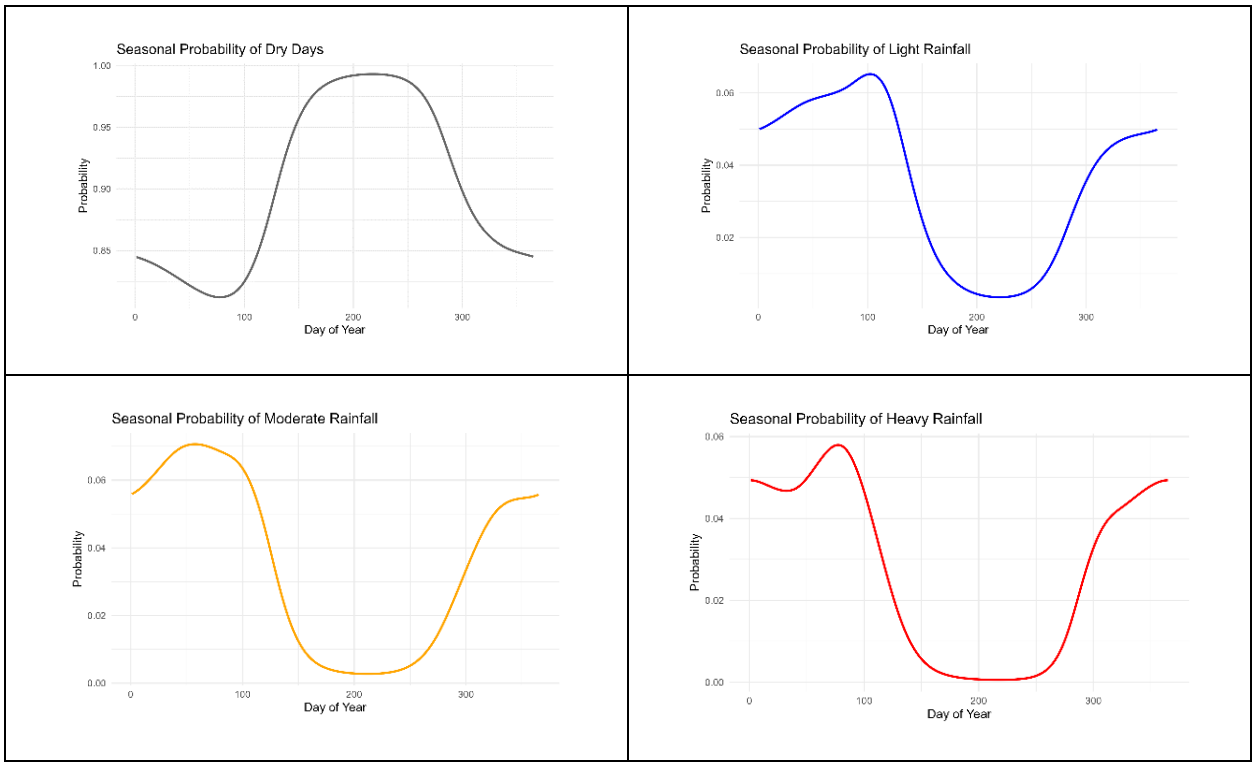


Figure 6. Seasonal probability splines for rainfall categories: Dry, Light, Moderate, and Heavy derived from the Extended GAM.

Figure 6 of the seasonal analysis of rainfall in Mosul shows that it is subject to a severe seasonal cycle and that the GAM model is able to model these cycles and show good classifications, and confirms the idea of using Extended GAM models in predicting complex climatic phenomena after the failure of CM to capture seasonality.

Conclusion

Based on the results and discussions, the highly erratic nature of daily rainfall in Mosul's semi-arid climate, characterized by a predominance of dry days and frequent extreme events, renders CM models unsuitable for probabilistic and classificatory forecasting, as they fail to account for seasonal variations. In contrast, the improved GAM model demonstrated its high flexibility in simulating nonlinearity and seasonal variations in rainfall. It proved that using time lags with seasonal time memory, along with weighted categories and adjusted decision thresholds, is the most reliable and effective option for classifying rare rainfall events and overcoming data imbalances. From this, it can be concluded that the significant hydrological similarity between light to moderate rainfall days and dry days limits the ability of time lags to accurately differentiate between them. Furthermore, the improved model, which combines time lags with weighted categories and adjustable decision thresholds, can be adopted as an effective tool in early warning systems for heavy rainfall and floods, and it produces better classifications of rainfall categories compared to traditional methods.

Acknowledgement

The work is self-sponsored by the authors under the supervision of the corresponding author.

Funding information

No funding.

Conflict of interest

The authors declare that there is no conflict of interest, financial or otherwise, related to the publication of this research paper.

Author contributions

Prof. Dr. Muthanna Subhi Sulaiman contributed to the conceptualization of the research, provided theoretical guidance on the climatological and geographical interpretations, and participated in critically reviewing and editing the final manuscript. Prof. Dr. Osamah Basheer Shuku developed the R programming codes, performed the data processing, statistical modeling (Markov Chain and GAM), and evaluation, conducted the formal analysis, wrote the primary draft of the manuscript, and acted as the corresponding author. All authors have read and approved the final version of the manuscript.

Ethics

Not applicable.

AI tool usage declaration

The authors declare that No AI-assisted tools were used and No AI system was used to generate original scientific claims, data, or research results.

Data availability

The daily rainfall dataset analyzed during the current study was obtained from the Iraqi Meteorological Organization and Seismological Monitoring (Climate Department) and was accessed through the doctoral dissertation and associated research by Nafea (2026) and Nafea & Al-Dujaili (2026). The dataset is not publicly available due to institutional restrictions but can be obtained from the corresponding author upon reasonable request.

References

Al-Nassar, A. R., Pelegrí, J. L., & Masciángioli, P. (2026). On the temporal variability of precipitation in Iraq: Arid-wet years and extreme events. *International Journal of Climatology, 46*(7), e70350. https://doi.org/10.1002/joc.70350

Anfossi, D., Cappa, C., Von Hardenberg, J., & Provenzale, A. (2002). Statistical analysis of three series of daily rainfall in north-western Italy. *Nuovo Cimento della Società Italiana di Fisica C, 25*(1), 99-122.

Chhatre, R. H., O'Donnell, J., & Ravishanker, N. (2026). Evaluating statistical methods for modeling and forecasting hourly precipitations. *Japanese Journal of Statistics and Data Science*, 1-32. https://doi.org/10.1007/s42081-026-00366-w

Chowdhury, D. P., Roy, D., & Saha, U. (2025). Study of rainfall occurrence process by Markov chain models and decision tree-based ensemble and boosting techniques. *Water Resources Management, 39*(6), 2857-2877. https://doi.org/10.1007/s11269-025-04095-9

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378. https://doi.org/10.1198/016214506000001437

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv*. https://doi.org/10.48550/arXiv.2008.05756

Hajani, E., & Sarma, G. (2023). Generation of rainfall data series by using the Markov chain model in three selected sites in the Kurdistan Region, Iraq. *AI in Civil Engineering, 2*(1), 5. https://doi.org/10.1007/s43503-023-00014-2

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning, 45*(2), 171-186. https://doi.org/10.1023/A:1010920819831

Hasan, S. T. H., & Shukur, O. B. (2025). Chaotic time series forecasting by using echo state network and autoregressive model. *Iraqi Statisticians Journal, 2*(Special Issue for ICSA2025), 1-11. https://doi.org/10.62933/faqd2g68

Hastie, T., & Tibshirani, R. (1986). [Generalized additive models]: Rejoinder. *Statistical Science, 1*(3), 314-318. https://doi.org/10.1214/ss/1177013604

Hyndman, R. J., & Grunwald, G. K. (2000). Applications: Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Australian & New Zealand Journal of Statistics, 42*(2), 145-158. https://doi.org/10.1111/1467-842X.00115

Nafea, K., & Al-Dujaili, A. M. J. (2026). *Rainfall prediction using machine learning in Nineveh Governorate, Iraq* [Conference paper]. Forum Geografi. https://doi.org/10.23917/forgeo.16759

Nafea, K. R. (2026). *Atlas of forecasting for climate Nineveh Governorate* [Unpublished doctoral dissertation]. Al-Iraqia University.

Ohaegbulem, A. F., & Nwobi, F. N. (2026). A stochastic assessment of monthly rainfall pattern: A case of Owerri, Nigeria. *Research in Statistics, 4*(1), 2655510. https://doi.org/10.1080/27684520.2026.2655510

Shukur, O. B. (2021). Using hybrid procedure in generalized additive model and classification tree to classify agriculture evaporation in Mosul city. *International Journal of Agricultural and Statistical Sciences, 17*(1), 1509- 1514. https://www.connectjournals.com/pages/articledetails/toc034635

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Vu, D. D., Nguyen, A. H., Nguyen, P. T., Thi, H. N., & Nguyen, T. T. N. (2025). Approaches to improving rainfall classification accuracy in Central Vietnam using machine learning models. *Korean Journal of Remote Sensing, 41*(2), 261-276. https://doi.org/10.7780/kjrs.2025.41.2.1.2

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548-1563. https://doi.org/10.1080/01621459.2016.1180986

Yusif, C. Y. (2024). *Statistical modeling of daily precipitation process in the context of climate change* [Doctoral dissertation, McGill University]. https://doi.org/10.82308/15181